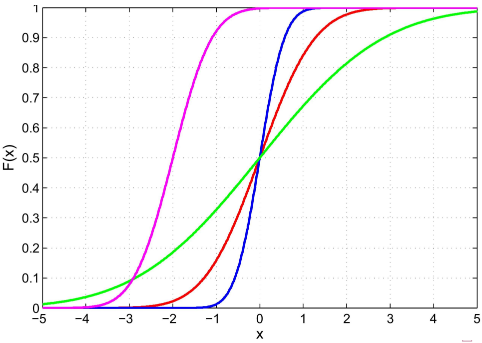
Прикладная статистика
1. Введение.

Родионов Игорь Владимирович
rodionov@bigdatateam.org

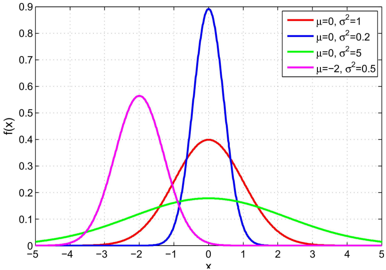
Случайная величина – некая функция от случая. Примеры: траты на телефонную связь за день, длительность разговора, время начала разговора.

Функция распределения случайной величины ξ :

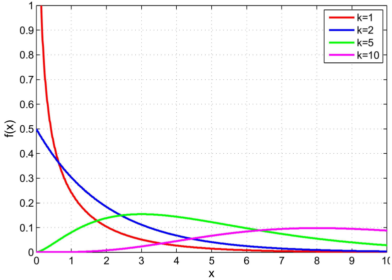
$$F(x) = P(\xi \leq x).$$



Плотность распределения $p(x) = F'(x)$, если производная есть.



$N(\mu, \sigma^2)$



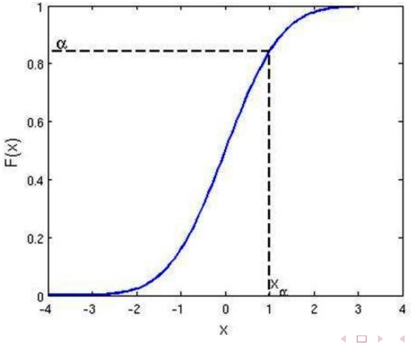
χ^2_k

Свойство плотности: $\int_{\mathbb{R}} p(x)dx = 1$.

Квантиль функции распределения F уровня α – такое x_α, что

$$F(x_\alpha) = \alpha.$$

Это означает, что с вероятностью α значение случайной величины ξ будет меньше x_α.



Характеристики распределений

Пусть случайная величина ξ имеет функцию распределения F. Тогда

- Математическое ожидание: $E\xi = \int_{\mathbb{R}} x dF(x)$;
- Дисперсия: $D\xi = E(\xi - E\xi)^2 = E\xi^2 - (E\xi)^2$;
- Коэффициент асимметрии (skewness)

$$Sk = \frac{E(\xi - E\xi)^3}{(D\xi)^{3/2}}$$

- Коэффициент эксцесса (excess, без вычитания 3 – kurtosis)

$$K = \frac{E(\xi - E\xi)^4}{(D\xi)^2} - 3$$

- Медиана: квантиль уровня $1/2$.

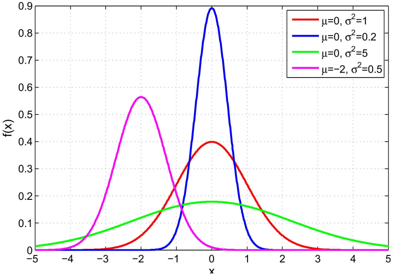
$\xi \sim N(\mu, \sigma^2)$, если плотность ξ равна

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Свойства:

1) $E\xi = \mu$; $D\xi = \sigma^2$;

2) Обозначим $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$, $\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy$, тогда $F_\xi(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.



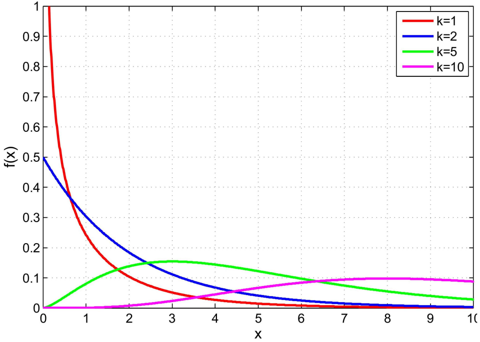
$X \sim \chi^2_k$, если плотность X равна

$$p_X(x) = \frac{x^{k/2-1}}{2^{k/2}\Gamma(k/2)} \exp(-x/2).$$

Свойства:

1) Если $\{\xi_i\}_{i=1}^k$ – нез.сл.в., $\forall i\ \xi_i \sim N(0,1)$, то $\sum_{i=1}^k \xi_i^2 \sim \chi^2_k$.

2) $EX = k$, $DX = 2k$.



Распределения Стьюдента и Фишера

1) Пусть $X \sim N(0,1)$, $Y \sim \chi^2_n$, X и Y независимы, тогда случайная величина

$$Z \stackrel{d}{=} \frac{X}{\sqrt{Y/n}}$$

будет иметь распределение Стьюдента с n степенями свободы, $Z \sim St(n)$ (также пишут $Z \sim T_n$).

2) Пусть $X \sim \chi^2_n$, $Y \sim \chi^2_m$, X и Y независимы, тогда случайная величина

$$Z \stackrel{d}{=} \frac{X/n}{Y/m}$$

будет иметь распределение Фишера с n и m степенями свободы, $Z \sim F(n,m)$.

Предельные теоремы

1) Закон больших чисел (ЗБЧ). Пусть $\{\xi_i\}_{i=1}^n$ – независимые одинаково распределенные случайные величины (н.о.р.сл.в.), $E|\xi_1| < \infty$, тогда $\forall \varepsilon > 0$ при $n \to \infty$

$$P\left(\left|\frac{\sum_{i=1}^n \xi_i}{n} - E\xi_1\right| > \varepsilon\right) \to 0.$$

2) Центральная предельная теорема (ЦПТ). Пусть $\{\xi_i\}_{i=1}^n$ – н.о.р.сл.в., $D\xi_1^2 < \infty$. тогда при $n \to \infty$

$$\frac{\sum_{i=1}^n \xi_i - nE\xi_1}{\sqrt{nD\xi_1}} \xrightarrow{d} N(0,1),$$

где $\xrightarrow{d}$ означает сходимость функций распределения.

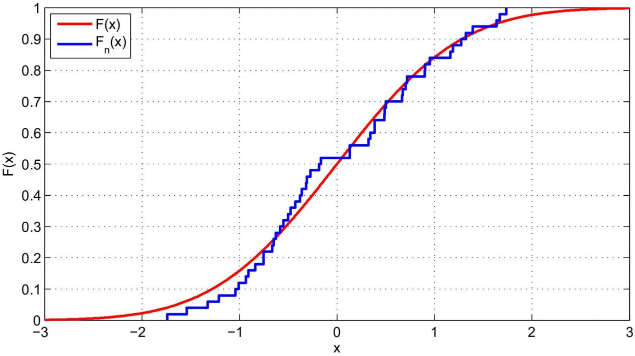
Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения P. Статистикой $T(X)$ называют функцию от выборки.
Примеры:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ – выборочное среднее;
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ – выборочная дисперсия;
- $\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ – выборочный k-тый момент;

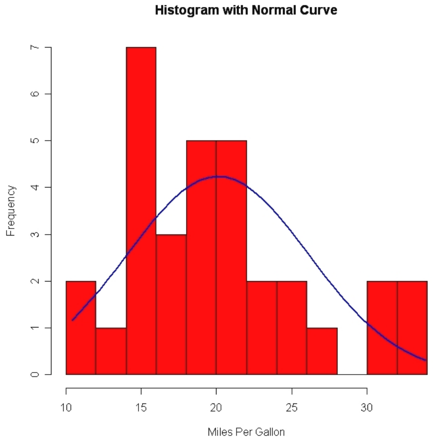
- Порядковые статистики $X_{(1)} = \min(X_1, \ldots, X_n)$, $X_{(2)} = \min(X_1, \ldots, X_n \backslash X_{(1)}), \ldots$ $X_{(n)} = \max(X_1, \ldots, X_n)$;
- Вариационный ряд: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$;
- Ранг элемента выборки (в вариационном ряду): $R(X_i) = r$, если $X_i = X_{(r)}$;
- Выборочная α-квантиль: $\widehat{Z}_\alpha = X_{([n\alpha])}$;
- Выборочная медиана:

$$\widehat{\mu} = \begin{cases} X_{(k+1)}, \text{ если } n = 2k+1; \\ \frac{1}{2}(X_{(k)} + X_{(k+1)}), \text{ если } n = 2k; \end{cases}$$

- $F_n(x) = \sum_{i=1}^{n} I(X_i \le x)$ – эмпирическая функция распределения.



- Гистограмма распределения:
$p_n(x) = \sum_{i,j} I(x \in B_j) I(X_i \in B_j)$, $\{B_j\}_{j=1}^k$ – разбиение.



Доверительные интервалы

Пара статистик $(T_1(X), T_2(X))$ называется доверительным интервалом для параметра θ уровня доверия $1-\alpha$, если $\forall \theta \in \Theta$

$$P_\theta(T_1(X) < \theta < T_2(X)) = 1-\alpha.$$

Последовательность пар статистик $(T_{1,n}(X), T_{2,n}(X))$ называется асимптотическим доверительным интервалом для параметра θ уровня доверия $1-\alpha$, если $\forall \theta \in \Theta$

$$P_\theta(T_{1,n}(X) < \theta < T_{2,n}(X)) \to 1-\alpha, \ \ n \to \infty.$$

Пример 1.
Пусть $X_1, \ldots, X_n$ – траты пользователя на мобильную связь за n дней. Построим доверительный интервал для средних трат EX_1 за день в предположении, что $X_1, \ldots, X_n$ независимы и одинаково распределены. Из ЦПТ,

$$\frac{\sum_{i=1}^{n} X_i - nEX_1}{\sqrt{nDX_1}} \xrightarrow{d} N(0, 1).$$

Вместо неизвестной DX_1 мы можем подставить выборочную дисперсию s^2, сходимость к нормальному закону сохранится.

Поскольку дробь $\frac{\sum_{i=1}^{n} X_i - nEX_1}{\sqrt{ns^2}}$ близка по распределению к нормальному закону, то

$$P\left(u_{\alpha/2} < \sqrt{n}\frac{\overline{X} - EX_1}{s} < u_{1-\alpha/2}\right) \approx 1 - \alpha,$$

где $u_{\alpha/2}$ и $u_{1-\alpha/2}$ – квантили $N(0,1)$. Решая эти неравенства относительно EX_1, получаем доверительный интервал уровня доверия $1-\alpha$

$$\overline{X} - u_{1-\alpha/2}\frac{s}{\sqrt{n}} < EX_1 < \overline{X} - u_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Пусть X – это выборка $(X_1, \ldots, X_n)$ из неизвестного распределения P_X. Основная задача – по выборке X сделать выводы о распределении P_X.

Предположим, что $P_X \in \mathcal{P}$, где $\mathcal{P}$ – некий класс распределений, которому заведомо принадлежит P_X.

Основы проверки гипотез

- **Основная гипотеза**: $H_0: \; P_X \in \mathcal{P}_0$, где $\mathcal{P}_0 \subset \mathcal{P}$.

  Пример. Пусть известно, что выборка имеет нормальное распределение, хотим проверить, что выборка распределена по закону $N(0,1)$. Тогда $\mathcal{P} = \{N(a,\sigma^2), a \in \mathbb{R}, \sigma^2 > 0\}$, а $\mathcal{P}_0 = \{N(0,1)\}$.

- **Альтернативная гипотеза (или альтернатива)**: $H_1: P_X \in \mathcal{P}_1$, где $\mathcal{P}_1 \subset \mathcal{P} \backslash \mathcal{P}_0$.

- Гипотеза называется **простой**, если $\mathcal{P}_0$ (или $\mathcal{P}_1$) состоит из одного распределения.

- **Статистика критерия**: $T(X)$ – такая статистика, что при $P_X \in \mathcal{P}_0$ мы либо знаем её распределение, либо можем оценить сверху вероятности её редких значений.

Основы проверки гипотез

- Если правило проверки гипотезы выглядит так:

$$\textit{если } T(X) \in S, \textit{ то отвергнуть } H_0,$$

  то S называется **критическим множеством**, а само правило называют **критерием**.

- Критерии бывают
  - двусторонние, $\{T(X) > u_{1-\alpha} \cup T(X) < u_\alpha\}$;
  - односторонние, которые делятся на правосторонние, $\{T(X) > u_{1-\alpha}\}$, и левосторонние, $\{T(X) < u_\alpha\}$;
  - более сложные.

Основы проверки гипотез

- **Уровень значимости** критерия: такое α, что $P_0(T(X) \in S) \leq \alpha \ \forall P_0 \in \mathcal{P}_0$.

- **Размером** критерия называется его минимальный уровень значимости, т.е. такое α, что

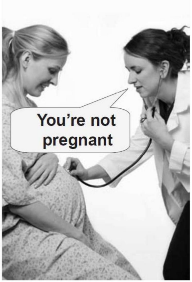
$$\alpha = \sup_{P_0 \in \mathcal{P}_0} P_0(T(X) \in S).$$

Уровень значимости выбирается исследователем. Его обычные значения – $0.1,\ 0.05$ или 0.01.

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода (False negative)
H_0 отвергается	Ошибка первого рода (False positive)	H_0 верно отвергнута

Type I error
(false positive)

Type II error
(false negative)



- **Функция мощности критерия**: $Q(S, P) = P(T(X) \in S)$.

Тогда $Q(S, P_0)$, $P_0 \in \mathcal{P}_0$, – вероятность ошибки I рода на распределении P_0, а $1 - Q(S, P_1)$, $P_1 \in \mathcal{P}_1$, – вероятность ошибки II рода на распределении P_1.

Получаем, что уменьшая вероятность ошибки I рода (т.е. уменьшая S), мы неизменно увеличиваем вероятность ошибки II рода, поэтому выбирать слишком низкий уровень значимости не рекомендуется.

- Критерий S **мощнее** критерия R, если уровни значимости этих критериев совпадают и $\forall P \in \mathcal{P}_1$ выполнено
$$P(T(X) \in S) \geq P(T(X) \in R).$$
- Критерий S называется **равномерно наиболее мощным** критерием (р.н.м.к.), если он мощнее любого другого критерия того же уровня значимости.

Р.н.м. критерии существуют далеко не во всех ситуациях. Пусть $X = (X_1, \ldots, X_n)$ – выборка размера n. В задаче различения двух простых гипотез $H_0 : P = P_0$ против $H_1 : P = P_1$ р.н.м.к. существует всегда, как утверждает следующая лемма.

Лемма Неймана-Пирсона.
Пусть $S_\lambda = \{x \in \mathbb{R}^n : \prod_i p_1(x_i) - \lambda \prod_i p_0(x_i) \geq 0\}$, где p_i – плотность распределения P_i по мере μ, $i = 0, 1$. Пусть критерий R того же уровня значимости, что и критерий S_λ, т.е. $P_0(X \in R) \leq P_0(X \in S_\lambda)$. Тогда
1) $P_1(X \in R) \leq P_1(X \in S_\lambda)$ (т.е. S_λ мощнее R);
2) $P_1(X \in S_\lambda) \geq P_0(X \in S_\lambda)$.

Пусть в Примере 1 про траты пользователя на мобильную связь мы хотим проверить гипотезу $H_0 : EX_1 = a$ против альтернативы $H_1 : EX_1 \neq a$. В полученный в Примере 1 доверительный интервал истинное EX_1 не попадает с маленькой вероятностью, поэтому если a не попало в доверительный интервал, то стоит предположить, что $EX_1 \neq a$. Получаем критерий

$$\text{если } \left| \frac{\sum_{i=1}^{n} X_i - na}{\sqrt{ns^2}} \right| > u_{1-\frac{\alpha}{2}}, \text{ то отвергать } H_0,$$

который называется **критерием Вальда**.

RML-тест

На практике часто возникает задача различения двух семейств распределений. Для этих целей можно использовать критерий отношения правдоподобия (RML-тест).

Пусть $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \in \Theta_1$, где $\Theta_0 \cap \Theta_1 = \varnothing$. Определим

$$f(X, \theta) = p_\theta(X_1) \cdot \ldots \cdot p_\theta(X_n).$$

Введем статистику

$$\lambda_n(X) = \frac{\sup_{\theta \in \Theta_0} f(X, \theta)}{\sup_{\theta \in \Theta_1} f(X, \theta)}.$$

Если $\lambda_n(X) < \lambda$, где λ определяется либо аналитически, либо моделированием, то отвергаем H_0.

В рамках байесовской парадигмы мы считаем, что на множестве распределений, из которых мы выбираем подходящее, задана некая вероятностная мера Q (априорное распределение) с плотностью q.

Пусть X – наблюдение из распределения $P \in \mathcal{P}$. Пусть семейство распределений $\mathcal{P}$ параметризовано, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, и $q(\theta)$ – априорная плотность на Θ. Формат вывода в рамках байесовского подхода таков:

$$q(\theta|X) = \frac{p(X|\theta)q(\theta)}{\int_\Theta p(X|\theta)q(\theta)d\theta},$$

где $p(X|\theta)$ – функция правдоподобия наблюдения X, а $q(\theta|X)$ – апостериорная плотность.

Байесовский подход

Пусть $H_0 : P \in \mathcal{P}_0 = \{P_\theta, \theta \in \Theta\}$ и $H_1 : P \in \mathcal{P}_1 = \{\widetilde{P}_\gamma, \gamma \in \Gamma\}$. Рассмотрим статистику

$$K = \frac{P(\mathcal{P}_0|X)}{P(\mathcal{P}_1|X)} = \frac{\int_\Theta f(X,\theta)q(\theta)d\theta}{\int_\Gamma \widetilde{f}(X,\gamma)\widetilde{q}(\gamma)d\gamma}.$$

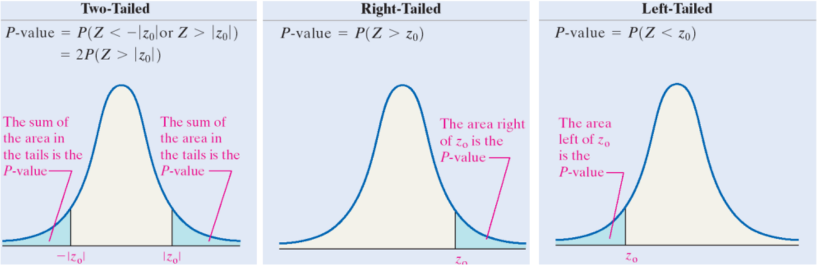
Распределение статистики K, как правило, найти тяжело, поэтому пользуются шкалой Джеффри:

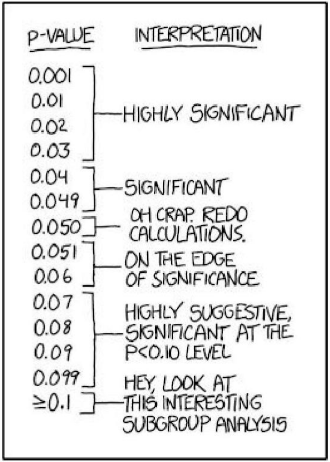
K	верна ли H_0?
1-3	нельзя определенно сказать
3-10	большие основания принять H_0
10-30	почти наверняка
>30	точно

Пусть значение статистики критерия $T(X)$ на наблюдении X равно t. Тогда р-значение – такая величина, которая является функцией от t и равна вероятности того, что $T(X)$ (на другой реализации наблюдения X) примет значение “экстремальнее”, чем t.

Правило проверки гипотезы с помощью р-значения выглядит так: если $p < \alpha$, где α – уровень значимости критерия, то отвергаем основную гипотезу.

В случае левостороннего критерия $p = P(T(X) < t)$, в случае правостороннего критерия $p = P(T(X) > t)$, в случае двустороннего критерия
$p = 2\min\{P(T(X) < t), P(T(X) > t)\}$.





Примерно так принимаются решения о значимости эффекта на основании p-значения (при основной гипотезе, что эффект незначим).

Пример (habrahabr.ru/company/stepic/blog/250527/). Допустим, мы хотим выяснить, существует ли связь между пристрастием к шутерам и агрессивностью у школьников. Для этого отобрали группу школьников, играющих в шутеры, и группу школьников, не играющих в компьютерные игры.

В качестве показателя агрессивности возьмём количество драк с участием конкретного школьника за месяц, в качестве основной гипотезы – что связи нет. Допустим, мы сравнили показатели 2 этих групп с помощью критерия хи-квадрат на уровне значимости 0.05 и получили p-значение, равное 0.04.

О чем говорит p-значение 0.04 в данном случае?

1. Компьютерные игры — причина агрессивного поведения с вероятностью 96%;
2. Вероятность того, что агрессивность и компьютерные игры не связаны, равна 0.04;
3. Если бы мы получили p-значение больше, чем 0.05, это означало бы, что агрессивность и компьютерные игры никак не связаны между собой;
4. Вероятность случайно получить такие различия равняется 0.04.
5. Ни один из вариантов не верен.

Ключевой вопрос: Допустим, при проверке некоторой гипотезы двумя критериями p-значение первого критерия оказалось меньше уровня значимости, а p-значение второго критерия больше. Как следует поступить: отвергнуть гипотезу или принять её?

Ключевой вопрос: Допустим, при проверке некоторой гипотезы двумя критериями p-значение первого критерия оказалось меньше уровня значимости, а p-значение второго критерия больше. Как следует поступить: отвергнуть гипотезу или принять её?

Ответ: Зависит от ситуации.

Спасибо за внимание!